The performance effect of feedback frequency and detail: Evidence from a field experiment in customer satisfaction[1]

March 2015

Pablo Casas-Arce
Arizona State University
casas.arce@gmail.edu

Sofia M. Lourenço
Universidade de Lisboa
slourenco@iseg.ulisboa.pt

F. Asís Martínez-Jerez
University of Notre Dame
asismartinez@nd.edu

Abstract: This paper presents the results from a field experiment that examines the effects of non-financial performance feedback on the behavior of professionals working for an insurance repair company. We vary the frequency (weekly and monthly) and the level of detail of the feedback that the 800 professionals receive. Contrary to what we would expect if these professionals were Bayesian and perfectly rational, more (and more frequent) information does not always help improve performance. In fact, we find that professionals achieve the best outcomes when they receive detailed but infrequent (monthly) feedback. The treatment group with frequent feedback, regardless of how detailed it is, performs no better than the control group (with monthly and aggregate information). The results are consistent with the information in the latest feedback report being most salient, and professionals in the weekly treatments overweighting their most recent performance, hampering their ability to learn.

[1] We thank Larbi Alaoui, Margaret Christ, Juanjo Ganuza, Gary Hecht, Chris Ittner, Michal Matejka, Andrei Shleifer, and seminar participants at the AAA Annual Meeting, IAP Cambridge Research Symposium, IE, MAS Midyear Meeting, Temple University Accounting Conference, Universitat Pompeu Fabra, University of Notre Dame, and University of Toledo for helpful comments. Sofia M. Lourenço gratefully acknowledges financial support from FCT—*Fundação para a Ciência e Tecnologia* (Portugal), research grant PTDC/EGE-GES/119607/2010 (national funding). All errors are our own.

I. Introduction

One of the main roles of performance measurement is to provide information for decision-making. Timeliness and comprehensiveness are usually regarded as desirable characteristics of information because they enable prompt and adequate responses to business threats and opportunities. However, the intensity of these attributes must be weighed against the decision maker's ability to process the relevant information. Too frequent information may result, for instance, in an overreaction to short-term factors, whereas too detailed information may cloud a decision maker's ability to identify general trends or issues. In this paper, we use a field experiment to analyze how the frequency and detail of performance feedback influences employee behavior in the context of customer satisfaction, a metric which is considered to be one of the most relevant indicators of the strategic health of a firm.

In collaboration with Multiasistencia—the leading Spanish business process outsourcer of repairs for insurance companies—we design and implement a field experiment in which we manipulate the non-financial performance feedback received by 800 home repair professionals (such as plumbers, masons, or painters) who work with the firm. We actively intervene in the feedback system by introducing a bonus that rewards the achievement of certain objectives in customer satisfaction as well as two process indicators. We vary the frequency of feedback information (weekly vs. monthly) and the level of detail included in the report (the average score of all jobs performed by a professional vs. the individual job scores for each professional). The field experiment design allows us to randomly assign professionals to different feedback regimes and analyze more cleanly the impact of the characteristics of interest.

If professionals were perfectly rational, they would use information efficiently to improve customer satisfaction. Therefore, more detailed and more frequent feedback should lead to better performance. However, more (and more frequent) customer satisfaction feedback does not always result in improved customer satisfaction scores. In fact, we find that professionals achieve higher scores when they receive detailed but infrequent (monthly) feedback. These results are consistent with the latest feedback report being most salient, and professionals overweighting the information contained in it. As a result, although detailed customer satisfaction feedback supplies information that helps professionals to improve the service they provide, feedback that is more frequent (and that consequently focuses on a shorter time horizon) ends up being less informative as previous information is disregarded in the face of new information.

Notably, we also show that the deterioration in performance of the professionals in the weekly treatments is explained by an irrational weighting of the most recent performance information and not by the rational abandonment of the effort to achieve the monthly bonus. There are two potential ways in which the rational abandonment of bonus targets could negatively impact the performance associated with frequent feedback. First, professionals in the weekly treatments could show worse performance because they learn earlier in the month that their performance disqualifies them from receiving the bonus or makes it very difficult to achieve, resulting in a rational abandonment of effort. Because professionals in the monthly treatments do not receive early performance updates, they do not have the option of abandoning the bonus target based on such feedback. To address this concern, we compare the performance of the different treatment groups during the first week of the month as a function of their performance in the last week of the previous month. We find that the weekly treatments perform worse than the monthly treatments in the first week if they had a negative report the previous week. This result cannot be explained by the rational decision to abandon the pursuit of the bonus, as bad performance the last week of the previous month has no impact on the chances of achieving the bonus in the current month.

A second concern may be that the performance in a given month is informative about the general difficulty of achieving the target and qualifying for a bonus, and therefore may affect the professional's decision to exert effort in future months. However, it is difficult to reconcile this possibility with the fact that only professionals in the weekly detailed treatment seem to conclude that the target is too difficult when they underperform in the last week of the month. Professionals in the monthly detailed treatment that underperform in the last week of the month have exactly the same information, but unlike the weekly detailed group, they do not show behavior consistent with giving up because they think the target is too difficult. If the professionals in the weekly detailed treatment infer that they are less likely to get the bonus, they must be overweighting the bad news from the previous week (their most recent performance report) relative to the professionals in the monthly detailed treatment.

These differences do not exist with respect to the process indicators included in the bonus system (e.g., the use of the Internet to schedule a service or finishing a repair on time). This is because the professionals receive immediate feedback simply by executing these tasks. Thus, the differences in the features of the formal feedback system do not result in any additional information, and do not affect professionals' knowledge about their performance or the way they process that information (Annett 1969). Consistent with this, we show that the operational performance is indeed the same in all four treatments.

This paper contributes to several streams of literature. First, Gennaioli and Shleifer (2010) argue that salience can help explain several of the behavioral biases in decision-making identified by psychologists. Our evidence shows that feedback in organizations can similarly induce behavioral responses that are not consistent with rationality, but that can easily be accounted for by assuming that feedback reports are salient. In particular, infrequent feedback increases the professional's ability to process information (especially if the feedback contains detailed information) and improves his or her decision-making.

In information economics, the contracting stream of research on performance measurement mainly focuses on how properties of information affect their inclusion in contracts (Feltham and Xie 1994; Prendergast 2002; Moers 2006) and occasionally how their inclusion in contracts affects business unit performance (Banker et al. 2000). The ultimate objective of this literature is to judge the strength of the performance metric in providing information to the *firm* about employees' choices (the control function). The design of performance metrics to facilitate the *employee*'s decision-making (the decision-making function) has been analyzed in the literature only rarely (Sprinkle 2003; Casas-Arce et al. 2014). In this study, we keep constant the incentive compensation baseline and show how changes in the detail and frequency of performance metrics affect decision makers' behavior.

Our work also contributes to the feedback literature by looking at the performance effects of the interaction of feedback frequency and feedback detail. Previous studies tended to examine these characteristics independently, with inconsistent results (e.g., Goodman et al. 2004; Chhokar and Wallin 1984). The presence of different moderators and the absence of an integrative theory of feedback are responsible for these not-always-well-understood inconsistencies (Kluger and DeNisi 1996). Northcraft, Schmidt and Ashford (2011) are an exception in the sense that they look at the joint effect of feedback frequency and detail. Their lab experiment asks subjects to perform four simple tasks simultaneously, with each task receiving a different feedback treatment. They study how the feedback characteristics affect the decision makers' allocation of resources among those tasks. In contrast, we use a field experiment to examine how detail and frequency affect decision makers' ability to process the relevant information to improve performance in the execution of a single job. We show that—contrary to what a model with perfectly rational decision makers would predict—more detailed and more frequent customer satisfaction information does not necessarily improve a professional's performance; in fact, the largest improvement in performance occurs in the group receiving more detailed but *less* frequent (monthly) feedback. Significantly, we observe these effects when feedback provides incremental knowledge of performance, but we do not observe them when the professional can also derive feedback instantaneously simply by executing the task (i.e., for process indicators such as whether a job is completed on time).

The structure of the paper is as follows. Section II reviews the relevant literature. Section III provides the institutional background of the research site, Multiasistencia, and the market in which it operates. The design of the field experiment is described in Section

IV. Section V provides motivation for the empirical tests. Section VI analyzes the empirical results, and Section VIII concludes.

II. Literature Review

Salience

Although economists often assume that people make rational inferences from all available information (Savage, 1954), psychologists have provided ample experimental evidence that is inconsistent with rational decision-making. For instance, Kahneman and Tversky (1972, 1974, 1983) show that people depart from Bayesian inference when processing information. Because individuals have limited cognitive resources, it would be too costly to process all available information; instead, they tend to overweight the data that is most salient (Taylor and Thompson, 1982).

Although many different biases in decision-making have been uncovered, recent work by Gennaioli and Shleifer (2010) and Bordalo, Gennaioli, and Shleifer (2012, 2013) shows that salience can account for a number of these behavioral anomalies and explain such behavior in a wide range of settings. In this paper, we provide further evidence that is consistent with the salience hypothesis, and we show how the reporting systems can be designed to mitigate the bias that results from salience. Specifically, we find that more frequent feedback, by directing attention to the most recent events, leads to worse decisions. Hence, reporting systems are most useful when providing detailed but infrequent information.

Feedback research

The traditional view in the literature is that feedback leads to performance improvement. In economic models of Bayesian updating, learning is a by-product of the utility maximization process in which the rational agent uses the new information provided by feedback to update her beliefs about the probable consequences of her choices and the impact on her utility (Savage 1954; Kiefer and Nyarko 1995). In the performance measurement and evaluation literature, feedback has a positive impact on performance because it improves learning and motivation (Ammons 1956; Ilgen et al. 1979; Kopelman 1986). However, a century-long body of research has shown that feedback does not uniformly improve performance (Balcazar et al. 1985; Kluger and DeNisi 1996; Alvero et al. 2001). There is now a consensus that the effect of feedback is contingent on the organizational setting in which it is provided and on the characteristics of the feedback itself (Balcazar et al. 1985; Kluger and DeNisi 1996). In particular, goal-setting and incentives stand out as features that appear to increase recipients' attention to feedback and improve the consistency of its effects (Locke and Latham 1990; Kluger and DeNisi 1996; Sprinkle 2000).

The specific feedback characteristics that researchers have looked at include, for example, the credibility and power of the source (Ilgen et al. 1979), whether the feedback is on individual or relative performance (Hannan et al. 2008), whether it is communicated privately to the recipient or made public (Hannan et al. 2008; Newman and Tafkov 2011), and whether it conveys a positive or negative message (Illies and Judge 2005). Two characteristics that have received special attention are the detail and frequency of feedback. The literature has long presumed that, in line with the Bayesian updating view, more detailed and more frequent feedback improves performance, although there are behavioral reasons for the excess of these characteristics to hamper the recipient's ability to process feedback information.

The traditional view of feedback detail is that an increase in detail improves performance. Thorndike's law of effect (1927) suggests that this is so because more detail permits a better identification of the behaviors that are reinforced and those that are punished. Detail also enhances the credibility of feedback, which becomes more believable when it is supported by specific examples (Leskcovec 1967). However, behavioral theories have questioned the positive effects of feedback detail. Very detailed feedback may direct the recipient's attention to specific events and result in the inappropriate generalization of a small number of salient situations rather than in a balanced learning inferred from all the information available, a phenomenon known as the law of small numbers (Tversky and Kahneman 1971; Rabin 2002). Moreover, when feedback provides very specific cues on how to improve performance, the recipient may disengage from the learning process, relying exclusively on the cues from feedback (Goodman et al. 2004).

Empirical evidence on the impact of feedback detail on performance is mixed: while some studies see a positive relationship, others do not, and some even find a U-shaped relationship between detail and performance (Goodman et al. 2004; Bilodeau 1969; Salmoni et al. 1984). This lack of consistency is caused in part by diversity in the definition of "detail," which can refer to traits as different as the level of precision of the feedback itself (Hannan et al. 2008) or the inclusion of advice on how to improve performance (Kim 1984). Also contributing to the lack of consistency are the different choices for the organizational design elements that interact with feedback, such as the incentive scheme (Northcraft et al. 2011; Hannan et al. 2008).

As in the case of feedback detail, the traditional view of feedback frequency in the literature is that more is better. From a learning standpoint, more frequent feedback allows the decision maker to revise her beliefs and try new strategies more often (Salmoni et al. 1984; Schmidt and Dolis 2009). From a motivational perspective, it contributes to the recipient's development of a sense of competence by allowing her to observe that her actions influence performance (Ilgen et al. 1979). Moreover, from an organizational point of view, an implicit value is given to metrics that are measured more frequently, which keeps the organization focused on those metrics (Reichheld 2006). However, behavioral theories argue that more frequent feedback may cause the recipient to lose perspective and pay more attention to the most recent performance. This orientation encourages a fire-fighting approach to problem solving rather than a long-term fundamental approach (Bohn 2000; Lurie and Swaminathan 2008). Additionally, more frequent feedback also increases the noise of the performance signal and could make it more difficult to learn (Bohn 1995; Lurie and Swaminathan 2008).

Although some experiments suggest that more frequent feedback may not improve performance (Chhokar and Wallin 1984, Lurie and Swaminathan 2008), most of the studies support the positive performance effects of frequent feedback (Kluger and DeNisi 1996; Balcazar et al. 1985; Alvero et al. 2001; Northcraft et al. 2011; Kang et al. 2005). A common explanation for the inconsistent results of these studies is that they suffer from methodological problems because they do not test purely for frequency but also add level of detail and/or other reinforcers such as training in the treatments.[2]

Previous studies have mainly looked at feedback frequency and detail independently (e.g., Goodman et al. 2004; Chhokar and Wallin 1984). Northcraft, Schmidt and Ashford (2011) are an exception in the sense that they look at the joint effect of both characteristics, but their lab experiment does not focus on how these characteristics affect the processing of information. Rather, they examine how the combination of feedback frequency and detail affects the salience of competing tasks and how decision makers allocate resources among those tasks. As expected, they find an additive effect.

Non-Financial Performance Measures (NFPMs) Research

In the mid-1990s, the management accounting literature expanded its scope to encompass the identification, measurement, and management of the drivers of strategic value creation (Ittner & Larcker 2001). Representative of this evolution was the emergence of "new" managerial accounting techniques such as the scoreboards of non-financial indicators (Kaplan & Norton 1996). The link with value creation was the ability of NFPMs to predict future financial performance. Although not always consistent, considerable evidence exists that these metrics may be leading indicators of financial performance (Amir and Lev 1996; Anderson, Fornell and Rust 1997; Ittner and Larcker 1998; Behn and Riley 1999; Banker et al. 2000). The evidence also suggests that the

[2] A parallel line of argument exists in the disclosure literature. Van Buskirk (2012) finds that more frequent disclosure leads to more speculation by investors. Bushee and Noe (2000) find that increases in a firm's disclosures—as measured by AIMR disclosure rankings—are associated with increases in speculative trading by institutional investors.

information content of NFPMs is affected by their attributes (Dikolli and Sedatole 2007; Chen, Martin and Merchant 2014).

Research in the area of NFPMs has predominantly focused on the preconditions for these metrics to improve incentive contracting (Feltham and Xie 1994; Prendergast 2002; Moers 2006) and occasionally on the effects of their inclusion in contracts (Banker, Potter and Srinivasan 2000). That is, the focus of the literature has been on the ability of the performance metric to provide information to the firm about employees' effort choices (*the control function*). The use of the performance metric to facilitate learning by the employee (*the decision-making function*), or how the different attributes of the NFPM impact employees' ability to process information, has been virtually ignored. Part of the reason for this oversight is the reliance on the agency notion of control, in which a performance measure is considered useless unless it provides information *about* the employee, rather than *to* the employee (Holmstrom 1979).

III. Research Setting

Multiasistencia is a business process outsourcing (BPO) firm that provides comprehensive claims management service for property and casualty insurance companies. The firm acts as the coordinator between clients with repair needs and a network of specialized home repair professionals. It is located in Europe and Latin America and is the industry leader in Spain, the country in which we base our study.

Multiasistencia's largest corporate clients in Spain are the insurance subsidiaries of major banks. The insurance companies hire Multiasistencia to manage the claims process for individual properties from the first report by the customer to the finishing touches of the repair.[3] Client relationships are governed by annual contracts. Typically, performance is formally reviewed on a monthly basis against service level agreements (SLAs) that include parameters of cost, timeliness, and quality of service. The CEO explained the nature of the interaction thus: “We assign a key account manager to each of the major insurance companies to oversee that client’s specific needs. The management team also maintains close connections with our largest corporate clients and communicates with their leaders approximately once a week.”

Multiasistencia employs over 300 customer service representatives (CSRs) in its call centers. There are separate phone banks for each of the four largest corporate clients and one general phone bank for overflow calls and calls from smaller customers. In a typical service intervention, the policyholder reports a claim by calling the insurance company, which redirects the call to Multiasistencia. The CSR at the call center makes an initial assessment of whether the caller’s claim is covered by the policy that he or she holds. Claims deemed to be covered by the policy are transferred to a regional dispatch office, where jobs are assigned to repair professionals as a function of the expertise required for the repair and the workload of the professional. Information from each call is recorded in a computer database. Throughout this process policyholders assume that they are interacting with the insurance company that has delegated its repair work to Multiasistencia.

Small repairs (less than three man-hours) are assigned to a repair professional who confirms or denies the coverage of the reported damage. If the professional confirms the

[3] We use the term “client” to refer to the insurance companies that outsource their repair work to Multiasistencia and the term “customer” to refer to the policyholder.

claim is covered by the policy, he or she carries out the repairs, completes a report, and closes the job in one visit. For larger jobs, the CSR assigns a professional to repair urgent damages and orders an assessment for the rest of the job. A claims inspector is sent to the site within two days of the call and issues a report to Multiasistencia. If the report justifies the claim, the professional is sent to complete the rest of the repairs. For repairs requiring more than one specialty (e.g., plumbing and glass repair), the intervention of each professional is scheduled sequentially by the dispatch center. Workflow and communications with and among professionals are managed and recorded through a system of hand-held devices (PDAs) supplied by Multiasistencia. At the end of each repair, a CSR contacts the policyholder to check that the repair has been completed.

The Repair Professionals

Multiasistencia works with a network of professionals. Repair professionals are not direct employees of Multiasistencia but are linked to the firm by relational contracts through which they receive a guaranteed stream of jobs. In exchange for receiving a guaranteed workflow, repair professionals commit to following Multiasistencia's operational procedures and giving priority to the firm's repairs.

Professionals are paid a fixed fee for visits that result in denial of coverage and for small jobs. For large jobs (those involving more than three hours of work), they are compensated on a variable scale based on the cost of materials and the number of hours needed to complete the repair. Small jobs account for 80% of all approved claims.

Prior to our experiment, there was no explicit incentive compensation system in place for repair professionals. However, Multiasistencia did track a set of operating indicators at the professional level.[4] Regional managers told professionals which indicators needed more of their attention, and better performers were implicitly rewarded with a heavier stream of work.

Customer Satisfaction

In 2012, Multiasistencia decided to make customer satisfaction a strategic priority. The CEO articulated it thus: "I want to take a qualitative leap in quality. I want to make it a differentiating factor. Today we are the best but we are not rewarded for that because the industry standard is a satisfied/not satisfied binary."

Contracts with insurance companies had traditionally specified target levels of customer satisfaction that were measured at the client level by surveying a sample of policyholders with repairs each month. The specific measure of customer satisfaction and the size of the surveyed sample varied from contract to contract. However, as the CEO noted at the time: "We do not have enough surveys to obtain a precise measure. If we could get a larger sample, and hence a more precise measure of each professional's performance in customer satisfaction, then we could give more weight to the outcome of satisfaction and less to the process metrics relative to what we are doing today." Thus, the firm decided to form a dedicated phone bank with CSRs who would perform the closing call for each repair and, at the same time, survey customer satisfaction.

Multiasistencia wanted a simple customer satisfaction metric that could be incorporated easily into a formulaic bonus plan. They decided to use a simplified version

[4] Some of the operating indicators followed were: repair time, use of the PDA to update the state of repair, percentage of customer complaint calls, and percentage of visits resulting in denial of coverage.

of the Net Promoter Score (NPS) metric.[5] The premise of NPS is that the best way to elicit a sincere and consistent response about the consumption experience is to ask customers whether they would refer the firm to others. The NPS creators believe that a customer makes a personal referral only when they believe the company offers a superior value and understands them. Thus, to assess the customer experience they ask: "On a scale of 0 to 10, how likely is it that you would recommend Company X to a friend or colleague? (0 = never; 10 = very likely)" (Reichheld 2003). Then, they classify customers as *promoters* (score 9–10) who loyally buy from the company and urge their friends to do so, *passives* (score 7–8) who are satisfied but unenthusiastic, and *detractors* (score 0–6) who would avoid any interaction with the company if they could. Multiasistencia decided to use the percentage of detractors among the customers surveyed in a month as the relevant metric for customer satisfaction.

To qualify for the bonus plan in any given month, a professional had to have zero customer complaints.[6] The bonus plan included three performance metrics: the number of detractors, the percentage of repairs fully scheduled with the PDA, and the percentage of repairs that ended in the standard time allotted for that type of job. Repair professionals received 0.70 euros per repair for each of the metrics in which their performance met or exceeded the respective targets. The targets were set by the management team and considered past performance of the different repair specialties.

These targets were:

- 100% of repairs fully scheduled with the PDA

[5] NPS is a trademark of Satmetrix Systems Inc., Bain & Co., and Frederick Reichheld.

[6] To count against a professional, the customer complaint had to be based on bad service quality; complaints about denial of coverage were excluded.

- 80% of repairs ended on time
- 0, 1, or 2 maximum detractors for professionals with less than 30, between 30 and 60, or more than 60 repairs in that month, respectively.

The customer satisfaction phone bank started to formally track customer satisfaction at the professional level in January 2013. Multiasistencia planned to use the data for the period January–March 2013 to help management understand the behavior of the metric. During this period, the information was shared across the management group but not with the repair professionals. In April, regional managers presented the detractors metric and the new bonus system to the repair professionals. The professionals learned about their performance for April via an email at the end of the month.

IV. Experimental Design

Our experiment immediately followed the events described above. Each of the professionals working for Multiasistencia was randomly allocated to one of four treatment groups that received different forms of feedback for a three-month period (May–July 2013). We manipulated two dimensions of that feedback: its frequency and level of detail. Professionals received feedback either on a monthly (M) or weekly (W) basis. Moreover, the feedback was either aggregate (A) or detailed (D). The combination of the two dimensions led to four treatments: MA, MD, WA, and WD. At the aggregate level, workers received only information about the total number of detractors during the reporting period. In the detailed treatments, workers received a list of the services with a detractor score (0–6) for the services they finished within the reporting period. The level of detail of the operating performance metrics did not change across treatments and professionals were informed of their percentage use of the PDA and percentage of services closed on time during the reporting period (week or month). Moreover, all professionals also received aggregate measures at the end of the month (as those were the basis for the bonuses they received). The four treatment groups are described in Figure 1.

The experiment began in late April when the company informed the professionals via e-mail of the new feedback protocol. This e-mail was tailored to the specific random assignment of each professional. Professionals were unaware that other types of feedback were provided to other individuals.[7] During the experiment, all professionals were informed about their performance according to their treatment condition.

After the initial information report at the end of April (the monthly aggregate report), which was common to all groups, those in the weekly information cycle received their first performance communication on May 6. Those in the monthly information cycle received their first performance communication on June 3. Subsequently, performance communications were issued on Mondays (for professionals in the weekly cycle every Monday, and for professionals in the monthly cycle on the first Monday after the end of the month).

For technical reasons, the company preferred to provide more timely information. Because of this, we were not allowed to have a balanced sample in all four treatments. Instead, 25% (75%) of the professionals received monthly (weekly) feedback, and 50%

[7] Professionals worked independently. Even jobs that required the input of multiple professionals (for instance, a broken pipe may have involved the work of both a plumber and a painter) did not require them to work simultaneously, and professionals rarely worked in the same location. Furthermore, the professionals were not unionized. For these reasons, information sharing among professionals was not common. We confirmed the lack of interaction among professionals in the pre-experiment survey. Although some sharing may still have occurred through informal networks, the short time frame of the experiment makes this possibility unlikely.

received aggregate or detailed performance information. Thus, we were left with about 100 professionals in treatments MA and MD, and 300 professionals in treatments WA and WD (see Table 1).

Because the company started monitoring customer satisfaction in January of 2013, we had four months of data available prior to the experiment.[8] Furthermore, we also administered a questionnaire one year prior to the beginning of the experiment to capture various characteristics of the professionals and to evaluate the risk of spillover across treatments inherent to an individual-level randomization.[9] We used this information to identify heterogeneous responses to the treatments. In addition, we observed four months of post-experiment performance. Figure 2 shows a detailed timeline of the field experiment.

V. Hypotheses

To understand the effects of feedback on performance, we develop a simple model that highlights the value of information for the different treatments. Suppose that the professional wishes to maximize the value of the services he provides $v(a_t, \theta)$, where $a_t \in A$ is an action taken by the professional in period t, and θ is an unknown parameter that determines the value of the different actions. v is the aggregate value of k individual services performed in the period, $s_{t,i}$ for $i = 1, \dots, k$. We take the period $t = (m, w)$ to correspond to a month m and a week w, and we assume four weeks in one month, i.e.

[8] Multiasistencia started computing this metric early in order to guarantee its consistency before introducing it into the incentive system.

[9] Because the pre-experiment survey was run so far in advance, we believe that it did not contaminate our results.

$w \in \{1,2,3,4\}$. If we denote by I_t the information available to the professional, then his objective at time t is

$$\max_{a_t \in A} E(v(a_t, \theta)|I_t)$$

The information available to the professional for making the decision depends on the feedback treatment $T \in \{MA, MD, WA, WD\}$. A Bayesian professional uses all available information in a rational way recalling all past feedback reports (Savage 1954). Therefore, such a professional under the MA treatment observes $I_{(m,w)}^{MA} = \{\bar{s}_n\}_{n<m}$, where $\bar{s}_n = \frac{1}{4k}\sum_{x,i} s_{(n,x),i}$ is the average performance for month n. A professional under MD treatment observes $I_{(m,w)}^{MD} = \{s_{(n,x),i}\}_{n<m,x,i}$. Under the WA treatment, he observes $I_{(m,w)}^{WA} = \{\bar{s}_{(n,x)}\}_{n<m,x} \cup \{\bar{s}_{(m,x)}\}_{x<w}$, where $\bar{s}_{(n,x)} = \frac{1}{k}\sum_i s_{(n,x),i}$ is the average performance for week x in month n. Finally, a professional under WD treatment observes $I_{(m,w)}^{WD} = I_{(m,w)}^{MD} \cup \{s_{(m,x),i}\}_{x<w,i}$.

If we assume that each realization of a service $s_{t,i}$ is informative about θ, then the information content of the different treatments is clearly ordered. Denote Blackwell's sufficiency order by $\succcurlyeq$. Then we have that $I_{(m,w)}^{MA} \preccurlyeq I_{(m,w)}^{MD}, I_{(m,w)}^{WA} \preccurlyeq I_{(m,w)}^{WD}$ for all m, w. Furthermore, notice that $I_{(m,1)}^{WA} \preccurlyeq I_{(m,1)}^{MD}$ because during the first week of the month, the MD treatment has more information than the WA treatment (they both have information about the same time periods, but the information is more detailed for the first treatment). Nonetheless, it is not possible to rank the information content of $I_{(m,w)}^{WA}$ and $I_{(m,w)}^{MD}$ for $w > 1$, as the WA treatment starts receiving further feedback about earlier weeks in month m while treatment MD does not. Similarly, notice that $I_{(m,1)}^{MD} \sim I_{(m,1)}^{WD}$, as both the MD and the WD treatments have signals that are equally informative during the first week of the month (they both observe detailed performance on all past services). But during the later weeks of the month, the WD treatment receives further updates, and hence has a more informative signal.

If we denote the expected performance of a professional under treatment T at time t by $Ev_t^T = \max_{a_t \in A} E(v(a_t, \theta)|I_t^T)$, then the following result follows directly from the ordering of the informativeness of the signals:

PROPOSITION 1. *The expected performance of a rational (Bayesian) professional satisfies:*

1. $Ev_t^{MA} \leq Ev_t^{MD}, Ev_t^{WA} \leq Ev_t^{WD}$ *for all* t.
2. $Ev_t^{WA} \leq Ev_t^{MD}$ *for* $t = (m, 1)$.
3. $Ev_t^{MD} = Ev_t^{WD}$ *for* $t = (m, 1)$.

The result shows that more information is always better for a rational professional, and therefore feedback is most effective when it is both detailed and frequent.

Suppose now that the professional is not perfectly rational. In particular, we will assume that the professional overweighs the last report when making inferences about the right course of action. In Gennaioli and Shleifer's (2010) terminology, the professional is a local thinker and the last feedback report is salient. To simplify matters, we will assume that the professional uses only the information contained in the last feedback report, disregarding all previous information. Hence, a local thinker professional observes $I_{(m,w)}^{L,MA} = \{\bar{s}_{m-1}\}$ under the MA treatment; $I_{(m,w)}^{L,MD} = \{s_{(m-1,x),i}\}_{x,i}$ under the MD treatment;

$I^{L,WA}_{(m,w)} = \{\bar{s}_{(m-1,4)}\}$ if $w = 1$ or $I^{L,WA}_{(m,w)} = \bar{s}_{(m,w-1)}$ if $w > 1$ under the WA treatment; and $I^{L,WD}_{(m,w)} = \{s_{(m-1,4),i}\}_i$ if $w = 1$ and $I^{L,WD}_{(m,w)} = \{s_{(m,w-1),i}\}_i$ if $w > 1$ under the WD treatment.

Because a local thinker disregards past feedback, the order of the signals based on their informativeness reverses. We now have $I^{WA}_{(m,w)} \preccurlyeq I^{MA}_{(m,w)}, I^{WD}_{(m,w)} \preccurlyeq I^{MD}_{(m,w)}$. At a given level of detail, the professional under the more frequent feedback disregards more information. As a result, the signal he uses is less informative. In fact, notice that $I^{WD}_{(m,w)} \preccurlyeq I^{MD}_{(m,w)}$ even for $w = 1$, despite the fact that, aggregating all past reports (as a Bayesian professional would do), both treatments have access to equally informative reports. As before, however, it is not possible to rank the signals that result from changing both the level of detail and the frequency. In this case, the WD treatment contains more detailed information, but over fewer services than the MA treatment. As a result, it is not possible to rank the two treatments.

The following result about expected performance for such professional follows:

PROPOSITION 2. *The expected performance of a local thinker professional satisfies:*

1. $Ev^{WA} \leq Ev^{MA}, Ev^{WD} \leq Ev^{MD}$ *for all* t.

The result highlights that more information is not always better when the professional is a local thinker. The way the information is presented affects the ability of the professional to process it, and in this case, we are likely to see the best results from feedback information under the detailed but infrequent feedback.

Along with the customer satisfaction metric, Multiasistencia provides feedback on two other process metrics: services finished on time and interventions scheduled through the PDA application. The professional does not know how the customer will rate the service event. In that sense, feedback on customer satisfaction provides new performance information. In contrast, performance in both of the process metrics is evident immediately, as the professional knows whether she schedules a job through the PDA or whether she finishes a job on time before she receives official feedback from the firm. Therefore, there is no new information in the feedback communicated to the professional for these metrics. If the signal is uninformative, $E(v(a_t, \theta)|I_t^T) = E\big(v(a_t, \theta)\big)$ and hence the professional can achieve the same expected performance under all treatments.

PROPOSITION 3. *If the feedback is uninformative, then the expected performance is the same for all treatments regardless of whether the professional is Bayesian or a local thinker.*

In the next section, we discuss how the data can shed light on the importance of these effects for the optimal release of feedback information.

VI. Experimental Results

In this section we compare the performance of professionals in the four treatments to identify the value of frequent and detailed feedback. The first result of the paper can be seen in Table 1, which provides summary statistics. It shows the average share of detractors for each of the four treatments. Professionals in all four treatments improve their performance (fewer detractors) between the pre-experiment and experiment periods, an effect that may be due to the introduction of the incentives, the introduction of the feedback, or a combination of both. Additionally, professionals' performance is similar across all treatments in the first four months of 2013, suggesting a successful randomization.

The three months of the experiment show the control group (MA) performing just as well as the weekly treatments (WD and WA), while professionals in the treatment MD show the most improvement in performance, achieving the lowest share of detractors of the four groups (8.37%).

A similar picture emerges when we look at the fraction of professionals with zero detractors in a month. This fraction increases for all groups during the experiment months, but it does so more markedly for treatment MD than for the others.

We also observe an improvement in the operational metrics included in the bonus program during the experimental period, but the improvement is very similar across all treatments.

We develop these insights below, with additional statistical analyses.

i. The effects of the amount and frequency of feedback

To formalize our inference about the treatment effects, we estimate various regression models. Because professionals are randomly assigned to one of the four treatments, we can estimate average treatment effects by comparing the average performance of the professionals assigned to each treatment during the three-month experimental period with the following regression:

$$y_{it} = \beta_0 + T_i\beta_1 + X_{it}\delta + \varepsilon_{it} \tag{1}$$

where y_{it} is the performance of professional i in period t, T is a vector of treatment indicators for each of the four treatments, and X is a vector of additional covariates. The controls in X include time effects, to control for time trends, and the repair specialty of the professional, to account for heterogeneity in professionals' characteristics. In the regressions, we drop the dummy for the control treatment (MA) so that the constant measures the average performance for this group and the coefficients on the other three treatment dummies measure the difference in performance relative to the control.

We begin by looking at the performance of professionals delivering customer satisfaction, as measured by the share of detractors. The estimates presented in column (1) of Table 2 show that the professionals in treatment MD perform better than those in the control group (MA). They manage to lower their share of detractors by 2 percentage points more than professionals in the MA treatment. This difference represents a sizeable 20% improvement relative to the 10% share of detractors in the control group. However, the professionals in the two weekly treatments (WA and WD) show no difference in performance with respect to the control group. The same results follow when we control in column (2) for month effects and for the specialty of the professional.

Because we also observe the professionals for the four months prior to the experiment, we compare the improvement in performance between the three months of the experiment and the previous four months for the four treatments using a difference-in-differences estimation. In this way we control for any heterogeneity across treatment groups that could have arisen spuriously during the random assignment process. We do so by including the vector of treatment indicators T, a dummy D indicating the treatment period, and their interaction in the following linear model:

$$y_{it} = \beta_0 + T_i\beta_1 + D_t\beta_2 + D_tT_i\beta_3 + X_{it}\delta + \varepsilon_{it} \qquad (2)$$

where the vector of covariates X now includes not only time and specialty effects, but also individual fixed effects to control for any unobserved heterogeneity. As before, we also drop the dummy for the control treatment, so that the interaction terms capture the performance of the other three treatments relative to the control group.

The estimates in model (3) of Table 2 show the basic difference-in-differences estimation, without any controls. We can see that the overall share of detractors is lower in the three months of the experiment than in earlier months, showing that performance improves after the introduction of the feedback system. Moreover, the professionals in treatment MD improve performance by more than the control group (MA). They manage to lower their share of detractors by 3.4 percentage points more than the 2.5 percentage points drop observed in the control group (representing a 46% and a 19% improvement, respectively, relative to the baseline 13% of detractors). The professionals in the two weekly treatments (WA and WD) also have fewer detractors, but their performance is not statistically different from those in the control group.

Adding the pre-treatment period also allows us to control for unobserved heterogeneity using individual (professional) fixed effects. Column (4) reports the results with monthly and individual effects. Again, we obtain the same results. Professionals in treatment MD improve their performance relative to the control group, but the weekly treatments are indistinguishable from that group.

Because we have a large number of observations with zero detractors (see Table 1), we also estimate a Tobit model in columns (5) to (8). In this case, however, we do not have individual effects because the maximum likelihood estimator of the Tobit model is inconsistent under fixed effects. We use specialty effects instead.

As expected, the coefficients from the Tobit model are larger (in absolute terms) than those of the linear probability model (OLS). Nonetheless, we find the same results. Professionals in the MD treatment perform significantly better than those in the control group, while the performance of professionals in the monthly treatments (WA and WD) and the control are indistinguishable.

Next, we turn to two alternative measures of customer satisfaction: the proportion of promoters and the average survey score. Because the proportion of observations with extreme values (0 or 100% of promoters, and 0 or 10 score) is very small, we only report the OLS results.[10] Columns (9) and (10) provide the results for promoters, controlling for month and professional fixed effects. Although the share of promoters increases over time, we find no differential effect for any of the treatments. Because the number of promoters does not affect professionals' compensation—but the number of detractors does—the professionals probably concentrate their efforts on using feedback to improve their performance in the most difficult services (the ones that were likely to yield them a low value in the survey).

If we compare the average score in the customer satisfaction survey (columns (11) and (12)), we again find that treatment MD is the only one that improves upon the control group. However, because the improvement in performance only happens for a fraction of the services provided by this group, the economic effect is smaller than in the results

[10] The results from the Tobit model are essentially the same, and are available from the authors upon request.

described above. The average score is 8.0 in the first four months of 2013. This score increases by almost half a point (or about 6%) during the experimental period for the control group, and increases by an additional 0.2 points (or 2.5%) for professionals in treatment MD.

The results presented in this section suggest that providing more detailed feedback is useful for improving performance. However, that is only the case when feedback is provided sparsely. Detailed feedback loses its usefulness when provided very frequently. In fact, the F-test shows that the effect of MD is significantly different from that of WD, suggesting that performance deteriorates when detailed information is provided more frequently. Similarly, providing more frequent feedback, even when it is less detailed, does not seem to help professionals improve their performance.

Taken together, the results suggest that professionals fail to process the additional information rationally. The recipient of frequent feedback may fixate on the most recent information, leading him or her to underweight or ignore evidence that is more distant in time and thus limiting the amount of information actually used in decision-making. This leads professionals to make the wrong inferences, reducing their learning and hampering performance improvement. By providing detailed but less frequent feedback, Multiasistencia communicates richer information in a single report, allowing professionals to identify true trends and ignore noise in the metric.

ii. Feedback on customer satisfaction vs. operational performance

We now turn to the effects of feedback on the two measures of operational performance that are also part of the incentive scheme. These two measures are of a very different nature than customer satisfaction: they capture the input of the professionals, while customer satisfaction is a measure of their output. Professionals can perfectly observe the performance of the former directly but they do not observe the latter until they receive feedback from the firm. Because the feedback does not provide any additional information on the operational performance measures, we should not expect to find differential effects for the different feedback treatments. The only possible exception would be if the feedback acts as a reminder that those dimensions of performance are important for the firm's management.

Table 3 estimates analogous models to those in Table 1 using OLS, where the dependent variable is either the share of services closed on time or the share of services that were properly recorded using the PDA.[11] As expected, the results show no differences among the treatments on these two dimensions. Not only are the coefficients statistically insignificant, but the magnitudes of the effects are also economically negligible.

iii. Salient feedback vs. dynamic incentives

The evidence presented so far is consistent with the hypothesis that more information about output-based performance measures is useful when provided within a timeframe that allows enough information to accumulate that professionals can make meaningful inferences from it. The same information becomes less useful when it is provided too frequently, as past feedback is disregarded.

[11] The share of observations with extreme values (0 or 100% of services) is very low for both measures. For this reason, we do not show the estimates from the Tobit model, although the results remain the same, and are available upon request.

In this section we provide further evidence that is consistent with professionals being local thinkers by disaggregating performance at the weekly level. We provide direct evidence that is consistent with professionals overreacting to bad news when feedback is frequent (weekly), and we also show that the evidence cannot be explained by dynamic incentive considerations.

Table 4 provides Tobit estimates of treatment effects using weekly data. Column (1) simply estimates average treatment effects controlling for specialty and time (weekly) effects. As with column (8) of Table 2, we find that professionals in monthly treatment MD improve their performance relative to the control group, while the weekly treatments WA and WD show no improvement (the coefficient on MD is almost statistically significant with a p-value of 0.11).

Next, we separate the treatment effects for the first week (before the weekly treatments receive any feedback about the current month) and for the rest of the month by estimating the following model:

$$y_{it} = \beta_0 + T_i\beta_1 + D_t^e\beta_2 + D_t^l\beta_3 + D_t^eT_i\beta_e + D_t^lT_i\beta_l + X_{it}\delta + \varepsilon_{it} \qquad (3)$$

where *t* now denotes weeks rather than months, D^e takes a value of 1 for the early part of the treatment months (the first week of each month) and 0 otherwise, and D^l takes a value of 1 for the later part of the treatment months. Because we omit the dummy for treatment MA, the coefficients on the other three treatments show their performance during the period relative to the control group.

The results in column (2) show that treatments WA and WD do just as well as the control group in the later part of the month, while they seem to perform worse in the first week (the coefficient is statistically significant and of similar size for both WA and WD). Furthermore, the weekly treatments do worse than treatment MD both in the first week and in the later part of the month. This result is inconsistent with rationality (proposition 1). If professionals were rationally abandoning the pursuit of their bonus targets after receiving bad news, we would expect to observe statistically indistinguishable performance across all treatments in the early weeks of the month and deteriorated performance in the later weeks of the month for the weekly treatments.

We can further disentangle whether the improper processing of information arises when the feedback report contains good news, bad news, or both. To do this, we split the effect of the second part of the month for those professionals with at least one detractor in the first week of the month from those with none. Model (3) in Table 4 estimates the following regression:

$$y_{it} = \beta_0 + T_i\beta_1 + D_t^e\beta_2 + D_t^l NoDetr_{it}\beta_3 + D_t^l Detr_{it}\beta_4 + D_t^e T_i\beta_e + D_t^l NoDetr_{it} T_i\beta_{l,NoDetr} + D_t^l Detr_{it} T_i \beta_{l,Detr} + X_{it}\delta + \varepsilon_{it} \qquad (4)$$

where *NoDetr* is an indicator function that takes a value of 1 if the professional does not receive any detractors in the first week of the month (and hence still qualifies for the bonus), while *Detr* indicates that there is at least one detractor in the first week.

The results in column (3) show that professionals in treatments WA and WD who receive at least one detractor in the first week of the month (and learn about it through their weekly feedback) perform significantly worse in the second part of the month than professionals in the control group who also receive a detractor the first week (but are unaware of it until the end of the month). If a professional in the WA and WD treatments receives no detractor in the first week, then his or her performance in the following weeks is indistinguishable from that of the professionals in the control group who also receive no detractor in the first week. By contrast, professionals in treatment MD do not perform worse than the control in the second part of the month if they receive a detractor in the first week, and they outperform the control group in the second part of the month if they do not receive a detractor in the first week.

Thus, professionals in the weekly treatments under-perform after receiving bad news. Notice, however, that this evidence could be consistent with the presence of dynamic incentives.[12] Because the bonus is paid monthly, the professionals in weekly treatments learn their interim performance, and can adjust their effort based on that information. These professionals may still process the information efficiently, but may fail to deliver higher performance because they learn that they do not qualify for a bonus well before the end of the month. In fact, if this is the case they may rationally abandon their pursuit of the bonus and lower their effort in the final weeks.

If professionals in weekly treatments are responding to dynamic incentives, we should observe, as we do, a drop in performance in the later part of the month if they receive a detractor early on, and an improvement otherwise. However, dynamic considerations should only affect performance at the end of the month. In contrast, if professionals are overreacting to frequent feedback information, we will see a drop in

[12] Non-linear incentive schemes are known to create dynamic incentives, with varied responses over time based on past performance (see, for instance, Casas-Arce and Martínez-Jerez 2009).

performance after they learn of a bad outcome regardless of the point in time. The differential predicted effect of each hypothesis should be at a maximum between the last week of a given month and the first week of the following month. Because the performance measure is reset each month for bonus calculation purposes, we should observe no difference in performance in the first week of the month based on performance the week before if the results for weekly treatments are driven by dynamic incentives. However, we would still see an effect if those professionals are overreacting to frequent feedback information.

We next separate the treatment effect for the first week of the month for those who receive at least one detractor in the preceding week (the last week of the previous month) and those who do not by estimating the following model:

$$y_{it} = \beta_0 + T_i\beta_1 + D_t^e NoDetr_{it}\beta_2 + D_t^e Detr_{it}\beta_3 + D_t^l\beta_4 + D_t^e NoDetr_{it}T_i\beta_{e,NoDetr} + D_t^e Detr_{it}T_i\beta_{e,Detr} + D_t^l T_i\beta_l + X_{it}\delta + \varepsilon_{it} \qquad (5)$$

where *NoDetr* now indicates that there is no detractor in the last week of the previous month, while *Detr* indicates that there is a detractor.

The estimates are in column (4) of Table 4. They show that the same results we find for the later part of the month also arise in the first week (if anything, they are even stronger). Most striking is the fact that the WD treatment performs worse than the MD in the first week of the month, despite the fact that both groups have the same amount of information and even when we compare only those who had the same performance the previous week. This evidence strongly suggests that the last feedback report is most salient for professionals. As a result, those in the weekly treatments overweight the importance of any detractors in the previous week, ignoring previous signals, and resulting in worse performance.

iv. Heterogeneous treatment effects

In this section we return to the monthly observations to find out whether there are heterogeneous responses to the treatment effects we identify. We focus on the level of ability or experience of the professional.

We use several measures to estimate how the treatment effects vary with the ability of the professional. First, we measure ability as the average number of detractors for each professional during the first four months of the year, before the experiment takes place. Column (1) shows the estimates of the treatment effects for the subsample of above-median-ability professionals (those with a below-median proportion of detractors in the pre-experiment period). The estimates for below-median-ability professionals are in column (2). The results show essentially identical effects for treatment MD in both subsamples, suggesting no heterogeneous treatment effects. In column (3), we can further see that the treatment has similar effects for all professionals. This column estimates the model with all observations and interacts the treatment effects with our ability measure (pre-experiment performance). Relative to professionals of the same ability level in the control group, those in treatment MD lower their share of detractors by 5.4% on average, and the improvement in performance does not seem to vary with initial ability. Hence, the benefits of detailed monthly feedback seem to be shared among professionals at all starting levels of performance.

We repeat the same analysis using the level of education of the professional as the measure of ability. Because this measure comes from the survey we administered before the experiment, the sample size is limited by the response rate and the turnover among professionals. Again, column (4) shows no evidence of heterogeneous effects.

Finally, we repeat the analysis using tenure at the firm as the measure of ability. Like education, this metric is self-reported in the pre-experiment survey. In column (5), the measure of high ability is a dummy indicating that the professional has been with Multiasistencia for longer than the average professional. Again, we do not find significant differences in the treatment effects for long- and short-tenure professionals.

These results show that the effects of the MD treatment are widespread and fairly homogeneous. The response is the same regardless of the professional's prior performance, level of education, or tenure at the firm. All professionals seem to benefit from detailed but not very frequent feedback.

v. Post-experiment performance

Finally, we look at the post-experiment performance of the professionals in the different treatments. After the three months of the experiment and in view of its results, Multiasistencia decided to provide all professionals with monthly and detailed detractor information. Once all the professionals receive feedback with the same level of detail and the same frequency, we expect to observe similar patterns of performance throughout the firm. We collect information on the four months following the experimental phase (August to November) to provide some additional robustness tests.

Table 6 presents the results. In column (1) we estimate the same Tobit model as in column (8) of Table 2 (difference-in-differences with specialty and time effects), but augmented with a dummy for the post-experiment period interacted with the treatment groups. (We do not report the treatment effects during the experimental months, as they are analogous to those in Table 2.) The results show that the improved performance with respect to the pre-experiment period persists in the post-experiment period, but now there are no differences among the treatment groups.

Column (2) then extends the Tobit model in column (4) of Table 4 to look at the overreaction to detractors (as in column (1), we do not report the coefficients for the experimental period, which are analogous to those in Table 4). The results show that as soon as professionals in the WA and WD treatments stop receiving weekly information, their overreaction to information about detractor(s) in the last week of the month goes away.

These results provide further evidence suggesting that the experiment results were not the product of chance. As soon as professionals stop receiving weekly information, their performance improves, and the deterioration of performance after receiving a bad report disappears.[13] In this regard, the fact that the MD treatment loses its advantage relative to the other treatments suggests that the effects of information are short-lived. This result is consistent with the assumption that professionals disregard past information in light of the latest feedback report, which is most salient to them.

[13] Incidentally, the post-experimental results also show that the professionals in the weekly treatments (the WD treatment in particular) did not withhold performance in the short term (for instance, by engaging in more experimentation to increase learning) in order to improve their performance in the long term.

VII. Conclusion

This paper presents evidence on how the characteristics of NFPMs drive improvements in performance by decision-making employees. Using a field experiment that manipulates the frequency and detail of the non-financial performance feedback received by professionals in a property repair company, we find that, in our setting, detailed information leads to a significant improvement in performance. However, contrary to what we would expect if professionals were perfectly rational, detailed information is only useful when provided sufficiently sparsely. When feedback is too frequent, professionals overreact to that information, and they perform significantly worse after receiving bad news than a control group with aggregate and less frequent information.

This evidence is consistent with decision makers (the repair professionals) not being able to properly process detailed information when it is provided too frequently. Professionals seem to fixate on the information contained in the last feedback report, disregarding past reports.

Our results are relevant for managers designing feedback systems. Advances in technology have facilitated the capture and prompt delivery of performance information within the corporation. In contrast with the common assumption that more and more frequent feedback always yields better results, our findings suggest that managers should weigh the benefits of detailed, immediately available feedback against the ability of the recipients to properly process that information.

One design feature of the feedback system in our study that should be generalized with special caution is the relevant frequency range. In our setting, monthly is the natural base for feedback frequency because of the nature of the tasks (finished in a few days and repeated several times a day) and the compensation cycle of the industry. Weekly measurement is thus a logical increase in frequency. In settings with longer or less frequently repeated tasks, the relevant feedback frequency options may differ.

Despite its limitations, the field experiment methodology has the significant advantage of allowing us to test the impact of alternative information and control system designs in the context where the final system will be implemented. In this sense, our paper contributes to the emerging body of field experiments in the management and economics literature (e.g., Levitt and List 2009).

References

Alvero, A. M., Bucklin, B. R., and Austin, J. (2001). An objective review of the effectiveness and essential characteristics of performance feedback in organizational settings (1985-1998). Journal of Organizational Behavior Management, 21(1), 3–29.

Ammons, R. B. (1956). Effects of knowledge of performance: A survey and tentative theoretical formulation. Journal of General Psychology, 54, 279–299.

Amir, E., and Lev, B. (1996). Value-relevance of nonfinancial information: The wireless communication industry. Journal of Accounting and Economics, 22, 3-30.

Anderson, E. W., Fornell, C., and Rust, R.T. (1997). Customer satisfaction, productivity, and profitability: differences between goods and services. Marketing Science, 16, 129–145.

Annett, J. (1969). Feedback and human behaviour. Harmondsworth, Middlesex, England: Penguin Books.

Balcazar, F., Hopkins, B. L., and Suarez, Y. (1985). A critical, objective review of performance feedback. Journal of Organizational Behavior Management, 7, 65–89.

Banker, R. D., Potter, G., and Srinivasan, D. (2000). An empirical investigation of an incentive plan that includes nonfinancial performance measures. Accounting Review, 75, 65–92.

Behn, B., and Riley, R. (1999). Using nonfinancial information to predict financial performance: the case of the U.S. airline industry. Journal of Accounting, Auditing, and Finance, 14(1), 29–56.

Bilodeau, E. A. (1969). Supplementary feedback and instructions. In Bilodeau, E. A. (Ed.), Principles of Skill Acquisition (pp. 235–253). New York: Academic Press.

Bohn, R. (1995). Noise and learning in semiconductor manufacturing. Management Science, 41(1), 31–42.

Bohn, R. (2000). Stop fighting fires. Harvard Business Review, 78(4), 82–91.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2012). Salience theory of choice under risk. The Quarterly Journal of Economics, 127(3), 1243–1285.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2013). Salience and consumer choice. The Journal of Political Economy, 121(5), 803–843.

Bushee, B., and Noe C. (2000). Corporate Disclosure Practices, Institutional Investors, and Stock Return Volatility. Journal of Accounting Research, 38: 171-202.

Casas-Arce, P., Martinez-Jerez, F. A. (2009). Relative performance compensation, contests, and dynamic incentives. Management Science, 55(8), 1306–1320.

Casas-Arce, P., F.A. Martinez-Jerez, and V.G. Narayanan. 2014. The Impact of Forward-Looking Metrics on Employee Decision-Making: The Case of Customer Lifetime Value. Working Paper University of Notre Dame.

Chen, C. X., Martin, M., and Merchant, K. A.. (2014). The effect of measurement timing on the information content of customer satisfaction measures. Management Accounting Research. Forthcoming.

Chhokar, J. S., and Wallin, J. A. (1984). A field study of the effect of feedback frequency on performance. Journal of Applied Psychology, 69(3), 524–530.

Dikolli, S. S., and Sedatole, K. L. (2007). Improvements in the information content of nonfinancial forward-looking performance measures: a taxonomy and empirical application. Journal of Management Accounting Research, 19, 71–104.

Feltham, G., and Xie, J. (1994). Performance measure congruity and diversity in multi-task principal/agent relations. The Accounting Review, 69(3), 429–453.

Gennaioli, N., and Shleifer, A. (2010). What comes to mind. The Quarterly Journal of Economics, 125(4), 1399–1433.

Goodman, J. S., Hendricks, M., and Wood, R. E. (2004). Feedback specificity, exploration, and learning. Journal of Applied Psychology, 89(2), 248–262.

Hannan, R. L., Krishnan, R., and Newman, A. H. (2008). The effects of disseminating relative performance feedback in tournament and individual performance compensation plans. The Accounting Review, 83(4), 893–913.

Holmstrom, B. 1979. Moral hazard and observability. Bell Journal of Economics 10 (1): 74–91.

Ilgen, D. R., Fisher, C. D., and Taylor, M. S. (1979). Consequences of individual feedback on behavior in organization. Journal of Applied Psychology, 64, 349–371.

Ilies, R., and Judge, T. A. (2005). Goal regulation across time: The effects of feedback and affect. Journal of Applied Psychology, 90, 453–467.

Ittner, C. D., and Larcker, D. F. (1998). Are non-financial measures leading indicators of financial performance? An analysis of customer satisfaction. Journal of Accounting Research, 36 (Supplement), 1–46.

Ittner, C. D., and Larcker, D. F. (2001). Assessing empirical research in managerial accounting: a value-based management perspective. Journal of Accounting and Economics, 32, 349–410.

Kahneman, D., and Tversky, A. (1972). Subjective probability: A judgement of representativeness. Cognitive Psychology, 3: 430-454.

Kahneman, D., and Tversky, A. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185: 1124-1131.

Kahneman, D., and Tversky, A. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgement. Psychological Review, 91: 293-315.

Kaplan, R.S., and Norton, D.P. (1996). The balanced scorecard: Translating strategy into action. Boston, MA: Harvard Business School Press.

Kang, K., Oah, S., and Dickinson, A. M. (2005). The relative effects of different frequencies of feedback on work performance: A simulation. Journal of Organizational Behavior Management, 23(4), 21–53.

Kiefer, N., and Nyarko, Y. (1995) "Savage-Bayesian models of economics," in "Essays in learning and rationality in economics and games" (eds.) A. Kirman and M. Salmon, Basil Blackell Press

Kim, J. S. (1984). Effect of behavior plus outcome goal setting and feedback on employee satisfaction and performance. Academy of Management Journal, 27, 139–149.

Kluger, A. N., and DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychological Bulletin, 119(2), 254–284.

Kopelman, R. (1986). Objective feedback. In Locke, E. A. (Ed.), Generalizing from the Laboratory to Field Settings (pp. 119-146). Lexington, MA: Lexington Books.

Leskovec, E. W. (1967). A guide for discussing the performance appraisal. Personnel Journal, 46, 150–152.

Levitt, S. D., and List, J. A. (2009). Field experiments in economics: The past, the present, and the future. European Economic Review, 53(1), 1–18.

Locke, E. A., and Latham, G. P. (1990). A Theory of Goal Setting and Task Performance. Englewood Cliffs, NJ: Prentice Hall.

Lourenço, S. (2014). Do Monetary Incentives, Feedback and Recognition matter for Performance? Evidence from a Field Experiment in a Retail Services Company. Working Paper, ISEG Universidade Técnica de Lisboa.

Lurie, N. H., and Swaminathan, J. M. (2008). Is timely information always better? The effect of feedback frequency on decision making. Organizational Behavior and Human Decision Processes, 108(2) (March), 315–329.

Moers, F. (2006). Performance measure properties and delegation. The Accounting Review, 81(4), 897–924.

Newman, A., and Tafkov, I. (2011). Relative Performance Information in Tournaments with Different Prize Structures. Available at SSRN 1973131.

Northcraft, G. B., Schmidt, A. M., and Ashford, S. J. (2011). Feedback and the rationing of time and effort among competing tasks. Journal of Applied Psychology, 96(5), 1076–1086.

Prendergast, C. (2002). The tenuous trade-off between risk and incentives. The Journal of Political Economy, 110, 1071–1102.

Rabin, M. (2002). Inference by believers in the law of small numbers. The Quarterly Journal of Economics, 117(3), 775–816.

Reichheld, F. F. (2003). The one number you need. Harvard Business Review, 81(12), 46–54.

Reichheld, F. F. (2006). The Ultimate Question: Driving Good Profits and True Growth. Boston, MA: Harvard Business School Press.

Salmoni, A. W., Schmidt, R. A., and Walter, C. B. (1984). Knowledge of results and motor learning: a review and critical reappraisal. Psychological Bulletin, 95(3), 355–386.

Savage, L. J. (1954). The Foundations of Statistics. New York, Wiley.

Schmidt, A. M., and Dolis, C. M. (2009). Something's got to give: The effects of dual-goal difficulty, goal progress, and expectancies on resource allocation. Journal of Applied Psychology, 94, 678–691.

Sprinkle, G. B. (2000). The effect of incentive contracts on learning and performance. The Accounting Review, 75(3), 299–326.

Sprinkle, G. B. (2003). Perspectives on experimental research in managerial accounting. Accounting, Organizations and Society, 28(2–3), 287–318.

Szymanski, D. M., and Henard, D. H. (2001). Customer satisfaction: a meta-analysis of the empirical evidence. Journal of the Academy of Marketing Science, 29, 16–35.

Taylor, S., and Thompson, S. (1982). Stalking the elusive vividness effect. Psychological Review, 89: 155-181.

Thorndike, E. L. (1927). The law of effect. American Journal of Psychology, 39, 212–222.

Tversky, A., and Kahneman, D. (1971). Belief in the law of small numbers. Psychological Bulletin, 76(2): 105-110.

Van Buskirk, A. (2012). Disclosure frequency and information asymmetry. Review of Quantitative Financial Analysis, 38: 411-440.

Figure 1—Experimental design

Treatment Group	Frequency of Feedback	Detail of Feedback
Group MA	Monthly	Aggregate
Group WA	Weekly	Aggregate
Group MD	Monthly	Detailed
Group WD	Weekly	Detailed

Figure 2—Timeline of the field experiment

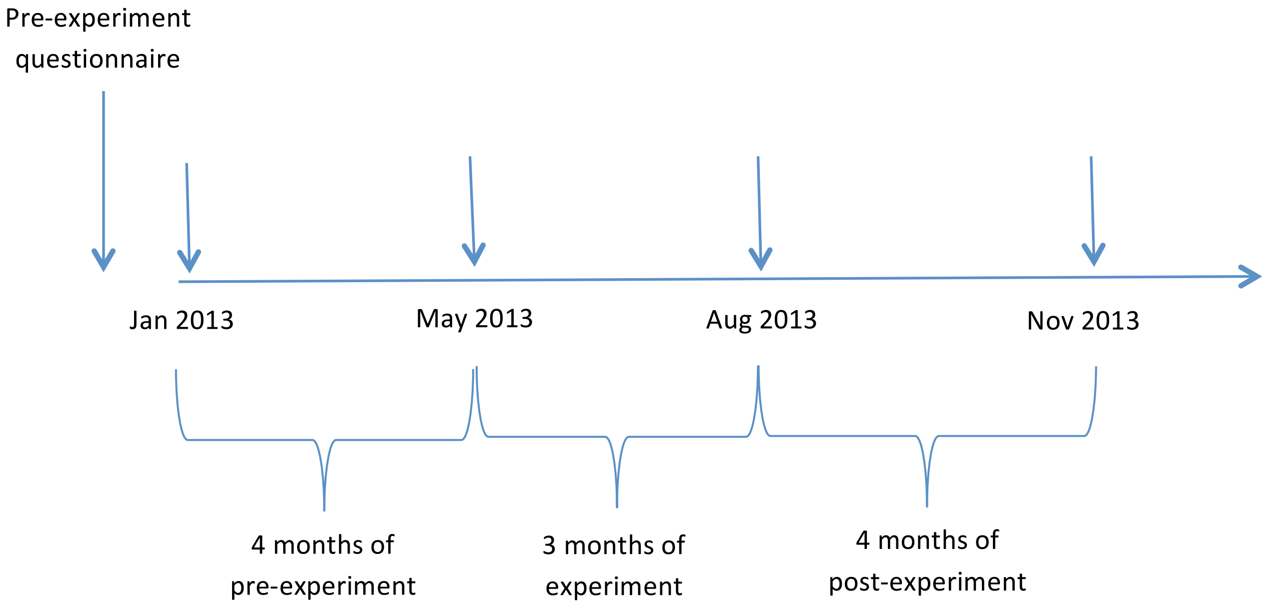
Pre-experiment questionnaire
Jan 2013
May 2013
Aug 2013
Nov 2013
4 months of pre-experiment
3 months of experiment
4 months of post-experiment

Table 1. Summary statistics

	January - April 2013			
	MA	**MD**	**WA**	**WD**
Detractors	13.02%	14.23%	14.44%	14.53%
No Detractors	32.13%	28.61%	28.91%	30.94%
On Time	50.11%	53.76%	50.02%	48.55%
PDA	76.50%	71.67%	76.39%	73.93%
Number of services per month	49.17	60.79	62.07	53.00
Number of surveys per month	15.21	19.48	19.93	17.50
Number of professionals	90	92	273	265

	May - July 2013			
	MA	**MD**	**WA**	**WD**
Detractors	10.53%	8.37%	10.52%	11.15%
No Detractors	39.10%	43.37%	38.05%	38.66%
On Time	54.26%	55.76%	53.64%	50.79%
PDA	79.30%	75.76%	77.40%	77.11%
Number of services per month	44.34	51.49	56.43	46.57
Number of surveys per month	12.39	15.07	16.78	14.52
Number of professionals	98	104	294	291

Notes: Detractors measures the proportion of services performed by a given professional with a score of 6/10 or lower. No Detractors measures the proportion of observations with zero detractors in a month. On Time measures the proportion of services closed in on time.

Table 2. The effects of feedback frequency and detail on customer satisfaction

	Detractor	Detractor	Detractor	Detractor	Detractor	Detractor	Detractor	Detractor	Promoter	Promoter	Score	Score
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
MD	-0.022*	-0.022**	0.012		-0.033*	-0.034*	0.021	0.019	0.004		-0.040	
	(0.011)	(0.011)	(0.012)		(0.019)	(0.018)	(0.016)	(0.015)	(0.017)		(0.078)	
WA	-0.000	-0.005	0.014		0.002	-0.008	0.022*	0.013	-0.003		-0.034	
	(0.010)	(0.010)	(0.009)		(0.015)	(0.015)	(0.013)	(0.013)	(0.015)		(0.063)	
WD	0.006	0.004	0.015		0.008	0.003	0.022*	0.017	-0.009		-0.060	
	(0.010)	(0.010)	(0.009)		(0.016)	(0.015)	(0.013)	(0.013)	(0.015)		(0.063)	
Experiment			-0.025**	-0.054***			-0.041**	-0.104***	0.056***	0.088***	0.262***	0.463***
			(0.012)	(0.013)			(0.017)	(0.019)	(0.019)	(0.020)	(0.075)	(0.083)
Experiment * MD			-0.034**	-0.039**			-0.053**	-0.053**	0.025	0.032	0.141	0.185*
			(0.016)	(0.016)			(0.024)	(0.023)	(0.026)	(0.023)	(0.109)	(0.101)
Experiment * WA			-0.014	-0.015			-0.020	-0.018	-0.004	0.004	-0.017	0.014
			(0.013)	(0.013)			(0.020)	(0.019)	(0.022)	(0.020)	(0.088)	(0.080)
Experiment * WD			-0.009	-0.011			-0.014	-0.013	0.003	0.005	0.021	0.050
			(0.014)	(0.013)			(0.020)	(0.019)	(0.022)	(0.020)	(0.088)	(0.080)
Constant	0.105***	0.093***	0.130***	0.162***	0.047***	0.338***	0.090***	0.332***	0.541***	0.506***	8.216***	8.000***
	(0.009)	(0.010)	(0.008)	(0.006)	(0.013)	(0.097)	(0.011)	(0.059)	(0.013)	(0.008)	(0.053)	(0.045)
Estimation	OLS	OLS	OLS	OLS	Tobit	Tobit	Tobit	Tobit	OLS	OLS	OLS	OLS
Time effects	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Specialty effects	No	Yes	No	No	No	Yes	No	Yes	No	No	No	No
Individual effects	No	No	No	Yes	No	No	No	No	No	Yes	No	Yes
R-squared	0.004	0.076	0.018	0.313	-	-	-	-	0.019	0.343	0.021	0.342
Observations	2,133	2,133	2,133	4,722	2,133	2,133	4,722	4,722	4,722	4,722	4,722	4,722

Notes: This table shows Tobit regressions of various measures of customer satisfaction captured at the monthly level: Detractor measures the proportion of services performed by a given professional with a score of 6/10 or lower; Promoter is the share of services with a score of 9 or 10/10; and Score is the average score over all the services with survey in the period. Experiment is a dummy variable that takes value of 1 during the three months of the experiment. MD, WA, and WD are treatment dummies that take a value of 1 or the professionals in the monthly-detailed, weekly-aggregate, and weekly-detailed treatments.

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 3. The effects of feedback frequency and detail on operational performance

	On Time	On Time	On Time	PDA	PDA	PDA
	(1)	(2)	(3)	(4)	(5)	(6)
MD	0.037**	0.036**	0.037**	-0.050***	-0.050***	-0.044**
	(0.017)	(0.017)	(0.016)	(0.019)	(0.019)	(0.018)
WA	-0.002	-0.002	0.013	-0.001	-0.001	-0.003
	(0.014)	(0.014)	(0.013)	(0.014)	(0.014)	(0.013)
WD	-0.016	-0.016	-0.005	-0.026*	-0.026*	-0.016
	(0.014)	(0.014)	(0.013)	(0.015)	(0.014)	(0.013)
Experiment	0.041**	0.087***	0.087***	0.029*	0.048**	0.054***
	(0.019)	(0.022)	(0.020)	(0.017)	(0.021)	(0.018)
Experiment * MD	-0.021	-0.021	-0.016	0.014	0.014	0.012
	(0.026)	(0.026)	(0.023)	(0.026)	(0.026)	(0.023)
Experiment * WA	-0.004	-0.004	-0.003	-0.018	-0.019	-0.018
	(0.021)	(0.021)	(0.019)	(0.020)	(0.020)	(0.018)
Experiment * WD	-0.020	-0.019	-0.021	0.004	0.004	0.000
	(0.022)	(0.021)	(0.019)	(0.020)	(0.020)	(0.018)
Constant	0.499***	0.456***	0.207***	0.763***	0.735***	0.867***
	(0.012)	(0.015)	(0.044)	(0.012)	(0.016)	(0.028)
Time effects	No	Yes	Yes	No	Yes	Yes
Specialty effects	No	No	Yes	No	No	Yes
Observations	4,962	4,962	4,962	4,745	4,745	4,745

Notes: This table shows Tobit regressions of two measures of operational performance captured at the monthly level: On Time measures the proportion of services closed in on time; PDA is the share of services scheduled with the PDA. Experiment is a dummy variable that takes value of 1 during the three months of the experiment. MD, WA, and WD are treatment dummies that take a value of 1 or the professionals in the monthly-detailed, weekly-aggregate, and weekly-detailed treatments.

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 4. Over-reaction to information vs dynamic incentives

	Detractor	Detractor	Detractor	Detractor
	(1)	(2)	(3)	(4)
Dummy I	-0.149***	-0.255***	-0.255***	0.388**
	(0.047)	(0.059)	(0.059)	(0.180)
Dummy I * MD	-0.053	0.004	0.004	0.104
	(0.034)	(0.063)	(0.063)	(0.108)
Dummy I * WA	0.030	0.093*	0.093*	0.196**
	(0.028)	(0.053)	(0.053)	(0.095)
Dummy I * WD	0.045	0.091*	0.091*	0.194**
	(0.028)	(0.054)	(0.054)	(0.096)
Dummy II				0.448***
				(0.172)
Dummy II * MD				-0.043
				(0.095)
Dummy II * WA				0.008
				(0.080)
Dummy II * WD				-0.029
				(0.081)
Dummy III		-0.137***	-0.171***	-0.133***
		(0.048)	(0.060)	(0.048)
Dummy III * MD		-0.068*	0.057	-0.066*
		(0.036)	(0.059)	(0.036)
Dummy III * WA		0.014	0.111**	0.010
		(0.030)	(0.049)	(0.030)
Dummy III * WD		0.034	0.127**	0.028
		(0.030)	(0.050)	(0.030)
Dummy IV			-0.123**	
			(0.049)	
Dummy IV * MD			-0.112***	
			(0.041)	
Dummy IV * WA			-0.022	
			(0.034)	
Dummy IV * WD			0.000	
			(0.034)	
Variable Definitions:				
Dummy I	Experiment	First week	First week	First week * Detr
Dummy II	-	-	-	First week * NoDetr
Dummy III	-	Later weeks	Later weeks * Detr	Later weeks
Dummy IV	-	-	Later weeks * NoDetr	-
Time effects	Yes	Yes	Yes	Yes
Specialty effects	Yes	Yes	Yes	Yes
Observations	17,372	17,372	17,372	17,372

Notes: This table shows Tobit regressions of a measure of customer satisfaction captured at the weekly level: Detractor measures the proportion of services performed by a given professional with a score of 6/10 or lower. Experiment is a dummy variable that takes value of 1 during the three months of the experiment. MD, WA, and WD are treatment dummies that take a value of 1 or the professionals in the monthly-detailed, weekly-aggregate, and weekly-detailed treatments. First Week is a dummy variable that takes value of 1 for the first week of every experiment month, while Later Weeks takes a value of 1 for the other weeks of the experiment months. Detr is a dummy variable that takes value of 1 if the professional had at least one detractor in the first week of the month (column 3) or in the last week of the previous month (column 4), and NoDetr takes value of 1 otherwise. We do not report the baseline coefficients on the MD, WA, and WD dummies for ease of presentation.
Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 5. Heterogeneous treatment effects

VARIABLES	Detractor (1)	Detractor (2)	Detractor (3)	Detractor (4)	Detractor (5)
Experiment	-0.187***	-0.051**	-0.099***	-0.063*	-0.069*
	(0.031)	(0.022)	(0.020)	(0.035)	(0.038)
Experiment * MD	-0.050	-0.050**	-0.054**	-0.104**	-0.088*
	(0.041)	(0.025)	(0.024)	(0.042)	(0.045)
Experiment * WA	-0.007	-0.022	-0.023	-0.036	-0.029
	(0.031)	(0.022)	(0.019)	(0.038)	(0.040)
Experiment * WD	-0.002	-0.012	-0.015	-0.051	-0.032
	(0.031)	(0.022)	(0.020)	(0.037)	(0.040)
Experiment * Ability			0.268	-0.012	0.015
			(0.216)	(0.055)	(0.043)
Experiment * Ability * MD			-0.026	0.038	-0.028
			(0.286)	(0.073)	(0.060)
Experiment * Ability * WA			-0.099	0.043	0.007
			(0.242)	(0.063)	(0.052)
Experiment * Ability * WD			0.102	0.006	-0.045
ment			(0.255)	(0.060)	(0.050)
Ability	Detractors	Detractors	Detractors	Education	Tenure
Sample	High ability	Low ability	All	Survey	Survey
Time effects	Yes	Yes	Yes	Yes	Yes
Specialty effects	Yes	Yes	Yes	Yes	Yes
Observations	1,772	2,950	4,700	1,372	1,394

Notes: This table shows Tobit regressions of a measure of customer satisfaction captured at the monthly level: Detractor measures the proportion of services performed by a given professional with a grade of 6/10 or lower. Experiment is a dummy variable that takes value of 1 during the three months of the experiment. MD, WA, and WD are treatment dummies that take a value of 1 or the professionals in the monthly-detailed, weekly-aggregate, and weekly-detailed treatments. We measure the ability of professionals with three measures: the average share of detractors for the months prior to the experiment (columns 1 to 3), self-reported measures of the level of education (column 4), and tenure at the firm (column 5). We do not report the baseline coefficients on the MD, WA, and WD dummies for ease of presentation.

Robust standard errors in pa

*** p<0.01, ** p<0.05, * p<0.1

Table 6. Post-treatment effects

	Detractor (1)	Detractor (2)
Post Period I	-0.049***	-0.013
	(0.018)	(0.119)
Post Period I * MD	-0.022	0.064
	(0.022)	(0.087)
Post Period I * WA	-0.006	0.047
	(0.018)	(0.072)
Post Period I * WD	-0.008	0.043
	(0.018)	(0.073)
Post Period II		-0.068
		(0.114)
Post Period II * MD		-0.096
		(0.082)
Post Period II * WA		0.025
		(0.063)
Post Period II * WD		0.006
		(0.065)
Post Period III		-0.014
		(0.043)
Post Period III * MD		0.004
		(0.033)
Post Period III * WA		-0.004
		(0.028)
Post Period III * WD		0.009
		(0.028)
Variable Definitions:		
Post Period I	Post-Experiment	First week * Detr
Post Period II	-	First week * NoDetr
Post Period III	-	Later weeks
Data	Monthly	Weekly
Time effects	Yes	Yes
Specialty effects	Yes	Yes
Observations	7,371	26,721

Notes: This table shows Tobit regressions of a measure of customer satisfaction: Detractor measures the proportion of services performed by a given professional with a grade of 6/10 or lower. Column 1 uses monthly data, and runs the same regression as in column 7 of table 2, with the additional variables shown here. Column 2 uses weekly data and runs the same regression as in column 4 of table 4, with the additional variables shown here. Post-Experiment is a dummy variable that takes value of 1 during the four months after the experiment (August to November). MD, WA, and WD are treatment dummies that take a value of 1 or the professionals in the monthly-detailed, weekly-aggregate, and weekly-detailed treatments. First Week Post is a dummy variable that takes value of 1 for the first week of every post-experiment month, while Later Weeks Post takes a value of 1 for the other weeks of the post-experiment months. Detr is a dummy variable that takes value of 1 if the professional had at least one detractor in the last week of the previous month , and NoDetr takes value of 1 otherwise.

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1